DOCUMENT RESUME

ED 408 342 TM 026 621

AUTHOR Cizek, Gregory J.; Husband, Timothy H.
TITLE A Monte Carlo Investigation of the Contrasting Groups Standard Setting Method.
PUB DATE Mar 97
NOTE 34p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, IL, March 24-28, 1997).
PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Cutting Scores; Educational Research; Educational Testing; Judges; *Monte Carlo Methods; *Pass Fail Grading; Sample Size; *Sampling; Simulation; Standards; *True Scores
IDENTIFIERS Angoff Methods; *Contrasting Groups Method; Experts; *Standard Setting

ABSTRACT

The contrasting groups method is one of many possible methods for setting passing scores. The most commonly used method is probably that developed by W. H. Angoff (1971), but it has been suggested that the Angoff method may not be appropriate for many standard setting applications in education. The contrasting groups method is explored as an alternative for educational research. To implement the contrasting groups method, experts are asked to make a dichotomous judgment about examinees, usually in the form of master/nonmaster, competent/not competent, certify/deny, and so on. All judged examines then take a test covering the content area of the domain of interest. This process results in two distributions of test scores, one for the group judged masters and one for the group judged nonmasters. These two distributions can be examined and used to derive a cutting score for the examination which is then applied to examinees who take the test for whom expert judgments of mastery are not available. In this study the contrasting groups graphing procedure was used in conjunction with various combinations of population and standard setting characteristics to examine the conditions under which it most reliably captures a known standard. A Monte Carlo approach was used to simulate and analyze populations with differing distributional forms, different percentages of master and nonmasters, various sample sizes, differing sampling strategies, and varying judge error rates. Overall, findings produced suggestions that the contrasting groups graphing procedure can be applied confidently to estimate a "true" cutting score in a variety of applications. Best sampling practices are discussed, and limitations of the expert criterion judgments are reviewed. (Contains 1 table, 8 figures, and 23 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
Gregory J. Cizek
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

ED 408 342

A Monte Carlo Investigation of the Contrasting Groups Standard Setting Method

March 1997

Gregory J. Cizek
Associate Professor of Educational Research and Measurement
350 Snyder Hall
University of Toledo
Toledo, OH 43606-3390
Phone: 419-530-2611
Email: gcizek@uoft02.utoledo.edu

Timothy H. Husband
Chair, Department of Mathematics
Siena Heights College
1247 E. Siena Heights Dr.
Adrian, MI 49221
Phone: 517-264-7647
Email: thusband@alpha.sienahts.edu

Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

BEST COPY AVAILABLE

TM026621

A Monte Carlo Investigation of the Contrasting Groups Standard Setting Method

Passing scores are used to mark two or more places on a score scale where important classifications or decisions are made. Some examples include: licensure or certification of competence for professional practice as in board examinations for physicians; credentialling, such as those awarded by the National Board for Professional Teaching Standards; or categorization, as in the National Assessment of Educational Progress (NAEP) achievement levels of basic, proficient, and advanced achievement. This study investigates one method of setting passing scores: the Contrasting Groups method.

Background

The Contrasting Groups method is only one of many possible methods for setting passing scores. In education, licensure, and certification, perhaps the most commonly used method was initiated by Angoff (1971). This method requires standard-setting participants to review test items and to provide estimations of the proportion of a subpopulation of examinees who would answer the items correctly. Angoff suggested that:

> A systematic procedure for deciding on the minimum raw scores for passing and honors might be developed as follows: keeping the hypothetical 'minimally acceptable person' in mind, one could go through the test item by item and decide whether such a person could answer correctly each item under consideration. If a score of one is given for each item answered correctly by the hypothetical person and a score of zero is given for each item answered incorrectly by that person, the sum of the item scores will equal the raw score earned by the "minimally acceptable person." (Angoff, 1971, pp. 514-515)

In practice, a footnoted variation to the procedure Angoff originally proposed has dominated applications of the method:

> A slight variation of this procedure is to ask each judge to state the probability that the 'minimally acceptable person' would answer each item correctly. In effect, judges would think of a number of minimally acceptable persons, instead of only one such person, and would estimate the proportion of minimally acceptable persons who would answer each item correctly. The sum of these probabilities would then represent the minimally acceptable score. (Angoff, 1971, p. 515).

In many applications of the procedure, the Angoff method is modified to facilitate less variable estimations. For example, many of the so-called "modified Angoff" approaches include two or more rounds of ratings. Also participants are often provided with normative data in one or more of the rounds of ratings, usually in the form of actual item difficulty indices.

Questions about the Angoff Method

Some researchers have suggested that the Angoff method may not be appropriate for many standard setting applications in education. For example, it has been suggested that making judgments about item content may be difficult for standard setting participants because it is a contrived task (Poggio, Glasnapp, & Eros, 1982).

More recently, Shepard, Glaser, Linn, and Bohrnstedt (1993) examined the use of the Angoff method to establish achievement levels for the NAEP and concluded that "the Angoff

method is fundamentally flawed for the setting of achievement levels (p. xxii).

Although their investigations only examined the use of the Angoff method, Shepard, Glaser, Linn, and Bohrnstedt (1993) also concluded that "other item-judgment methods are fundamentally flawed" (p. xxiv); that is, other commonly used methods such as the Ebel (1972) and Nedelsky (1954) approaches. Further, they were "skeptical" that the Angoff method would "be defensible in other contexts [other than the NAEP] (e.g., setting minimum standards based on all-or-none judgments about essential knowledge for a specific vocation" (p. xxiv).

A Search for Alternatives

Although rumors of the death of the Angoff method may be greatly exaggerated, it seems prudent to pursue investigations of alternatives. Jaeger (1989a, p. 492) has classified prevailing standard setting methods into two categories: test-centered continuum methods and examinee-centered continuum methods.

The Angoff, Ebel, and Nedelsky methods are classified as test-centered methods, because subjective expert[1] judgment is focused and exercised primarily upon test items. In examinee-centered methods, the focus of judgment is on examinees. Livingston and Zieky comment that "the main advantage of these [examinee-centered] methods is that people in our society are accustomed to judging other people's skills as adequate or inadequate for some purpose--especially in educational and occupational settings" (1982, p. 31).

One frequently recommended examinee-centered procedure is known as the Contrasting Groups method (see Livingston & Zieky, 1982 for a full description), also referred to as "an extension of the familiar known-groups validation procedure" (Berk, 1976,

p. 4). The Contrasting Groups method can be applied to traditional tests, such as those using multiple-choice or other selected-response formats. However, the method may be particularly useful for setting standards on complex, performance-based measures such as writing assessments, performances of a physical task, or other demonstrations in which the task for standard setting participants is simply to judge whether the performance exceeds some criterion. With the increasing use of performance assessments, the need to investigate alternatives to test-centered methods seems apparent.

The Contrasting Groups Method

To implement the Contrasting Groups method, experts are asked to make a dichotomous judgment about examinees, usually in the form of master/nonmaster, competent/not competent, certify/deny certification, and so on. All judged examinees then take a test covering the content area in the domain of interest. This process results in two distributions of test scores: one for the group judged to be masters, and another for the group judged to be nonmasters. These two distributions can then be examined and used to derive a cutting score for the examination which is then applied to examinees who take the test, but for whom expert judgments of mastery/non mastery are unavailable.

There are several solutions for deriving a cutting score from the two score distributions. For example, Livingston and Zieky (1982) illustrate a procedure in which a cumulative frequency distribution of all examinees' scores is plotted, showing the percent judged to be passing at each score point or interval. This distribution of the total group is then smoothed, using one of several possible methods. Livingston and Zieky observe that, to derive the final recommended passing score, "one logical choice is the test score for which

the 'smoothed' percent-qualified is exactly 50 percent" (p. 40).

Another variation of the Contrasting Groups method involves identifying a point that minimizes the overall impact of errors of classification. For example, the graphing method (used in this study) can be used in which the test score distribution of the group judged to be nonmasters is graphed on the same scale with the distribution of the group judged to be masters. Figure 1 illustrates a passing score obtained using the contrasting groups graphing method, with the cutting score indicated as C_x.

Insert Figure 1 about here.

Objectives

It is recognized that, in all standard setting, no "true" cutting score exists, except perhaps as the mean judgment of a population of all qualified participants in the standard setting process (Cizek, 1993). This mean judgment, however, can be conceptualized as a point on a score scale. Thus, if a point on a score scale can be thought of as a true passing score, then the ability of a standard setting method to capture that point can be evaluated.

In this study, the Contrasting Groups method was studied in conjunction with various combinations of population and standard setting characteristics in order to examine the conditions under which that method most reliably captures a known standard. Only the Contrasting Groups graphing method was examined in this study.

A Monte Carlo approach was used to address five specific research questions: 1) What effect do different shapes of parent distributions have on the ability of the Contrasting Groups

method to detect a "true" cut score?; 2) What effect do differing proportions of masters and nonmasters have on the passing score as determined by the Contrasting Groups method?; 3 and 4) How do the manner of sampling from the parent population and sample size affect the ability of the Contrasting Groups graphing procedure to estimate a cutting score?; and 5) What is the effect of various combinations of judge error rates, sampling strategies, and base rates on the accuracy of the Contrasting Groups graphing method?

Method

This study used a Monte Carlo design to simulate and analyze populations with differing distributional forms, different percentages of masters and nonmasters, various sample sizes, differing sampling strategies, and varying judge error rates. The goal of the analysis was to determine an optimum strategy for determining a cut score, using the contrasting groups method, given the various constraints and modifications modeled in the simulation (i.e., different parent population distributions, differing criterion levels for passing and judge classification error rate, distinct sampling strategy and contrasting sample sizes).

The steps that were followed to accomplish that goal are outlined in Figure 2, which shows a flow chart for this simulation. The process began by simulating five different populations of true scores (highly negatively skewed, moderately negatively skewed normal, moderately positively skewed, and highly positively skewed) on a scale from 0 to 100. Each distribution contained 10,200 cases; these distributions are depicted in Figure 3. The distributions were generated using the software package Minitab (Mathworks Inc., 1995). Because the proportion of masters in a population may have an effect on the resulting passing score when the contrasting groups method is used, this proportion was included as a studied

variable with three levels of true masters that would likely be close to those encountered in real populations (60%, 70%, 80%). These levels were applied to the populations to identify "true" cutting scores; i.e., the points P_{20}, P_{30}, and P_{40} were calculated and used as the true cutting scores.

Insert Figures 2 and 3 about here.

From the population distributions, subdistributions of "true masters" and "true nonmasters" were created. For example, when the symmetric population with a mean of 55 and standard deviation of 16.2 was used with the assumption of 80 percent masters in the population, the 20th percentile of the distribution (41.9) was used as the true passing score. Any score below 41.9 would be considered a true nonmaster (TNM) and any score greater than 41.9 would be considered a true master (TM). In practice, obviously, the Contrasting Groups method is not applied when "true" mastery or nonmastery status is known. Therefore it seemed reasonable to include an error term for each value in the subdistributions. To include this characteristic, a quantity was added to each score to simulate observed scores; the added value was a random number from a normal distribution with a mean of zero and a standard deviation equal to the standard error of measurement (SEM), which was computed using the formula for the SEM provided by Schaeafer, Carlson, and Matas (1986).

The five populations of 10,200 scores were then each divided into six equal intervals. The intervals were created to facilitate subsequent proportional sampling strategies. Populations with the appropriate specifications were constructed so that a fixed number of

scores fell in each of the six intervals to appropriately define the population. For example, the symmetric distribution has 400 (approximately 4%) in the first and sixth intervals, 1300 (12.7%) scores in the second and fifth intervals, and 3400 (33%) in the third and fourth intervals.

For sampling from the populations, it was decided to create a process that might reflect what occurs in actual implementations of the Contrasting Groups method. In practice, for example, a sample of examinations, such as student essays, are provided to a panel of judges who rate the performance as meeting or not meeting some criteria. In this study, the process was modeled by including various sample sizes of essays that might be used by judges (n_{essays} = 24, 102, 1020). Sampling from the parent population was performed with five different approaches that were hypothesized to have an effect on the precision of the derived cut score using the contrasting groups procedure (negatively skewed, uniform, symmetric, positively skewed, extreme groups). The samples were constructed by the following process. First, the abscissa of the population score distribution was partitioned into six equal intervals. These intervals were created to facilitate the following sampling plans:

1 - The first sampling from the population of scores was uniform, hence, an equal number of scores were randomly selected from each interval of the parent population.

2 - The second sampling was symmetric, with fewer scores randomly sampled from the tails and heavier sampling from the middle of the test score distribution (i.e., the distribution of scores across the six intervals in a symmetric distribution was in percentages of 4%, 12.7%, 33%, 33%, 12.7% and 4%). All symmetric samples (of sizes 24, 102, and 1020) had frequency distributions which were distributed in

approximately the same proportion. For example, for the symmetric samples of size 104 from the populations, 4 (4% x 104) scores were randomly chosen from the first and sixth interval, 13 from the second and fifth intervals, and 34 from the third and fourth intervals.

3 and 4 - The third and fourth sampling patterns were sampling in the appropriate tails of the distribution (in a fashion similar to the symmetric sampling).

5 - The fifth strategy employed extreme groups, involving sampling 50% from both tails of the distribution; i.e., from the first and last intervals.

The next step involved simulating judges' errors in classifying essays as either mastery or nonmastery. To simulate this error, uniform distributions were first created on the interval (0, 1). Judges were then, as a group, assigned an error rate of 10, 20, and 30%, and sampling proceeded. Thus, for example, when using the judge error rate of 30%, and a value from the uniform distribution less than or equal to .30 indicated that judges made an error in classifying the examinee; a value greater than .30 would indicate a correct classification. Furthermore, to represent situations that would most likely occur in practice, the test score interval was partitioned by values one standard deviation above and below the true cut score. This is consistent with the idea that judges' classification errors are unlikely for scores (essays) unusually far above or below the borderline of mastery/nonmastery. If the selected score was further than one standard deviation from the true cut score, no misclassification error was assigned.

This process resulted in two distributions of judged masters and nonmasters, which were used to derive estimated cutting scores. However, because such distributions usually

appear to be irregular in shape--especially when sample sizes are small--a smoothing strategy was implemented. The smoothing strategy used in this study was the locally weighted scatterplot smoothing (LOWESS) method suggested by Cleveland (1979, 1985). The smoothed y-value for any (x,y) point was accomplished by the following process:

1. Selecting a fraction, f, of all points, using the points closet in x-value on either side of the point under consideration. This selection is called f.n points. More points might be selected from one side of the interval that the other.

2. Calculating weights using the distance between each point in the selected fraction and the point to be smoothed as follows:

$$\text{weight} = [1-(\text{distance from selected point}/\text{maximum distance between selected point and the f.n points})^3]^3$$

This equation produces weights that have approximately a "normal" distribution in the neighborhood of the selected point (e.g., most of the weight applied near the point and very little in the end points of the interval under consideration).

3. Performing weighted linear regression on all points in the selected fraction of the data using weights from the process performed in step 2 (above) to produce an initial smoothed value.

4. Limiting the influence of outliers on the results on the above computations by doing two more iterations of step 3 (called the robust steps) with new weights calculated as follows:

$$\text{weight}=[1-(|\text{residual for points from previous step}|/(6)(\text{median of all }|\text{residuals}|\text{ from previous step})^2]^2$$

This strategy requires the user to decide, a priori, on the sampling fraction to be used in selecting points for the smoothing and to select a number of iterations. For this study, a

sampling fraction of .20 was used and iterations were set at 2. These decisions followed preliminary results which showed that a small sampling fraction was required when small samples were used (n=24), and that smoothing was improved if the sampling fraction was increased when larger samples were used. However, to maintain consistency in this portion of the contrasting groups procedure, the lower sampling fraction of 0.2 was used for all curve smoothing. Because this is a calculation intense process, a program was written for the software that performed the smoothing task and calculated the desired output using two iterations of the above described routine.

The operational cut score was defined as the point at which the master and nonmaster distributions intersected. Occasionally, the iterative smoothing process located multiple intersections of the criterion curves; therefore an algorithm was established which selected the score at which the maximum number of masters occurred, which corresponds closely to the intent of Livingston and Ziekey's recommendation (1982). An example of this occurrence and application of this rule of multiple intersections is shown in Figure 4. In Figure 4 an arrow identifies a cut line that the algorithm identifies as the location of the desired test standard (C_x).

Insert Figure 4 about here.

Results of the simulation were evaluated by calculating the following variables: C_x (the operational cutting score); p_o (the agreement coefficient); κ (the proportion of agreement corrected for chance) (Subkoviak, 1984); Ω (incremental validity, described by Berk, 1976);

and δ (a statistic used in this study to represent the average difference between the "true" cut score for a population and the operational cut score derived from application of the Monte Carlo procedures).

Ten iterations for each of the 675 comparisons, ([population (5) x [sample size (3) x sampling strategy (5) x [master population base rate (3)] x [judge error (3)]) were performed. The means, standard deviations, and ranges of p_o, κ, Ω, and δ were recorded. Graphs of the various combinations of these situations were constructed as well.

Results

Five specific research questions were addressed in this study. This section presents each research question and evidence from the data that bears upon it.

Research Question 1: What effect do different shapes of parent distributions have on the ability of the Contrasting Groups method to detect a "true" passing score?

One aspect of this study examined the effect of different shapes of parent distributions on the ability of the Contrasting Groups method to detect a "true" cut. To address this question the computed δs were compared. To guard against unusual samples that might bias results for comparison purposes (i.e., not representative of the parent population) sampling was done "like the parent population." It should be noted that one of the five sampling strategies applied to each population was to sample from the six intervals in proportion to the population's density over each interval.

For this question, large sample (n = 1020) comparisons using three population base rates over five different parent distributions, and judge error rate of 30%, revealed that the

Contrasting Groups graphing method identified a cutting score within $\pm 1SEM/\sqrt{10}$ of the "true" cutting score in all populations studied. Examination of the statistic δ revealed that it was smallest and of least variability around zero for negatively skewed parent populations. Figure 5 shows the results of this criterion (proximity to true cut score) as applied to evaluate how well the contrasting groups procedure performed under the specific manipulations of the parameters studied.

Insert Figure 5 about here.

Figure 6 shows box plots of the δs for each distribution over all base rates (80%, 70%, and 60%). As shown in the figure, the highly negatively skewed distribution had the smallest variability around zero. Furthermore, Figure 6 illustrates that in all of the populations the strategies tended to overestimate the true cut score, with the exception of the symmetric population, in which the true value was frequently underestimated.

Insert Figure 6 about here.

Generally, as base rates decreased, or as parent distributions become more negatively skewed, the contrasting groups procedure yielded more accurate results. However, a notable exception to this generalization was the highly positively skewed distribution.

Research Question 2: What effect do different proportions of masters and nonmasters have on the ability of the Contrasting Groups method to detect a "true" passing score?

The second research question addressed the effect of differing base rates (i.e, proportions of masters and nonmasters) on the passing score derived via the contrasting groups method. To examine this, the cut score consistency measure δ was calculated and compared across all sampling strategies and population base rates for each distribution in the study. A sample size of n = 1020 and a judge error rate of 30% were used to highlight and assist in stabilizing any pattern that might exist. No marked overall change in pattern in the sampling strategies' ability to identify the true cut over the three base rates was identified. With few exceptions, all distributions with δ values within ± 1 SEM/$\sqrt{10}$ at one base rate, tended to stay with ± 1 SEM/$\sqrt{10}$ at the other base rates with 10 iterations of the simulation.

The only dramatic exception to this result occurred when the extreme group sampling strategy was used. When sampling was done with this technique in distributions that were not positively skewed and when the base rate in the population was high (e.g., 80% masters), the result was a substantial overestimate of the true cut score. Figure 7 shows an example of this phenomenon.

Insert Figure 7 about here.

With the distributions and sampling strategies used, the extreme groups strategy was the only strategy for which this pattern was observed. In some instances, the estimate improved with decreases in base rate, most notably, when sampling in a negatively skewed

fashion from the distributions. In every instance the average δ decreased when the population base rate decreased from 80% to 60%.

Research Questions 3 & 4: What effects do sample size and sampling strategy have on the ability of the Contrasting Groups method to detect a "true" passing score?

The third and fourth questions addressed the effects of sample size and sampling strategies from the parent population on the ability of the contrasting groups graphing procedure to estimate a cutting score. Of the three sample sizes used in this study, the simulation and graphing procedure could not be completed for sample sizes of 24, because the estimation procedures utilized did not converge in 3-25% of the cases, depending on the particular combination of population and sampling strategy employed. However, results were obtained for the other sample sizes and sampling strategies.

Results of this analysis varied in nonsystematic ways that cannot be easily described. Results of all distributions followed similar general patterns; they failed to consistently capture the true cut score at the lowest sample size, had moderate success at the 102 sample size and comparatively better success at the 1020 sample size (again with the exception of the extreme group sampling strategy). For samples of size 1020, all δs were within $\pm 1SEM/\sqrt{10}$ for the uniform sampling strategy over all populations, all judge error rates and all population base rates. The negatively skewed sampling strategy had the same success except for the highly positively skewed population at the 30% judge error rate (δ = 1.96, $SEM/\sqrt{10}=1.23$).

Research Questions 5: What effect does mean judge error rate have on the ability of the

Contrasting Groups method to detect a "true" passing score?

Judge error rates are a concern in any decision making process. Although this simulation did not take into consideration the unique way a person's judgment may be flawed, it did attempt to make a reasonable estimate of how a panel of judges would, on an average, make errors. Of course, the more proficient the panel of judges, the smaller the group error rate. However, even if the panel made no errors in assessing the observed score papers they received (e.g., papers with scores above the true cut score were classified as masters, below the true cut-score as nonmasters), there would still be classification errors using the observed test score (e.g., a nonmaster may have guessed well yielding an observed score above the true cut score). It appears that these errors (measurement error and judge error) combine in idiosyncratic ways depending on the population base rate and the sampling strategy, which might affect judge opinion (e.g., when judges are asked to rate very poor papers or very good papers, they are less likely to error). The effect of various combinations of judge error rates, sampling strategies, and base rates on the accuracy of the Contrasting Groups graphing method was the target of the fifth research question.

This question was examined using only a moderately negatively skewed--a distribution shape that might be frequently assumed in proficiency or certification testing (Ziomek & Szymczuk, 1983). Comparisons of δs were made across the five sampling strategies and three judge error rates. Results indicated that, for a given sampling strategy, the judge error rate did not appear to have a substantial effect on the accuracy of the estimation of the "true" population cut score. For example, Figure 8 reveals the positively skewed, uniform, symmetric, and negatively skewed sampling strategies had δ values close to zero regardless of the judge error rate; while the extreme groups strategy always exceeded a δ of 1.59.

Insert Figure 8 about here.

To study the effect of judge error on derived cut score more closely, the standard deviations of the derived cut scores were examined for the situation of the moderately negatively-skewed population, sample size of 1020, over all sampling strategies, base rates and judge error rates. Table 1 gives the standard deviations of the derived cut scores for a moderately-negatively skewed population, sample size of 1020, over all sampling strategies, base rates and judge error rates. The table reveals steady or slight increases in variability with increases in base rate and judge error for all sampling strategies except for the extreme groups strategy which demonstrated large variability and erratic behavior at the lowest base rate level.

Insert Table 1 about here.

Limitations, Summary, and Discussion

The results reported in this paper are subject to several limitations. This investigation was limited to use study parameters within a small, but reasonable range of values. Also, this study was based upon simulated, as opposed to "real" data. Results presented in this paper represent only a portion of the full findings. For example, the major criterion of interest reported in this paper (δ) seemed most relevant and interpretable. Results and

analysis for other the criterion variables, p_o, κ, and Ω are forthcoming. Finally, this study utilized only the Contrasting Groups graphing method; although this is a commonly used method, there are other ways to derive a cutting score using the contrasting groups method which were not included in this study [e.g, decision-making accuracy approach (Berk, 1976), base rates analysis (Peters, 1981), utility function analysis (Overall & Klett, 1972), and discriminant function analysis (Koffler, 1980)]. The relative advantages of these alternatives have not received much attention in the literature.

Despite these limitations, we believe that this study provided some results that may be useful to practitioners and should provide suggested avenues for future research. Overall, our findings produced some evidence that the contrasting groups graphing procedure can confidently be applied to estimate a "true" cutting score in a variety of applications that resemble those often encountered in "real life" situations. In particular, we note four significant findings: 1) sample sizes of approximately 100 seem to be sufficient for the procedure to produce stable estimates; 2) negatively skewed and symmetric sampling strategies seem to provide the best results; 3) judge error rates--even fairly substantial error rates--did not appear to have a substantial effect on the accuracy of the estimation of the "true" population cut score; and 4) the accuracy of the procedure generally increases as population base rates for mastery decreased from 80% to 60%.

In conclusion, we also restate one concern about the contrasting groups method that has been mentioned by others who have studied this method: the validity and dependability of the criterion judgments. Because human judgments are involved, those judgments assigning examinees to "known" master or nonmaster groups are fallible. It is equally necessary to examine the adequacy of these classifications as it is to examine the psychometric

characteristics of the predictor (e.g., the examination). Like nearly all standards, initial classifications by experts of mastery and nonmastery to form known groups cannot be assumed to be "true" classifications.

REFERENCES

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), Educational measurement (pp. 508-600). Washington, DC: American Council on Education.

Berk, R. A. (1976). Determination of optional [sic] cutting scores in criterion-referenced measurement. Journal of Experimental Education, 15(3), 291-295.

Cizek, G. J. (1993). Reconsidering standards and criteria. Journal of Educational Measurement, 30(2), 93-106.

Cizek, G. J. (1996). Standard setting guidelines. Educational Measurement: Issues and Practice, 15(1), 13-21, 12.

Cizek, G. J. (1996). Passing scores. [NMCE Items Module]. Educational Measurement: Issues and Practice, 15(2), 20-31.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. Journal of the American Statistical Association, 74, 829-837.

Cleveland, W. S. (1985). The elements of graphing data. Monterey, CA: Wadsworth.

Ebel, R. L. (1972). Essentials of educational measurement. Englewood Cliffs, NJ: Prentice-Hall.

Jaeger, R. M. (1989a). Certification of student competence. In R. L. Linn (Ed.), Educational measurement, 3rd ed. (pp. 485-514). New York: Macmillan.

Jaeger, R. M. (1989b). Selection of judges for standard setting: What kinds? How many? Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Jaeger, R. M. (1991). Selection of judges for standard-setting. Educational Measurement: Issues and Practice, 10(2), 3-6, 10, 14.

Koffler, S. (1970). A comparison of approaches for setting proficiency standards. Trenton, NJ: New Jersey Department of Education. (ERIC Document Reproduction Service No. ED 181 028)

Livingston, S. A., & Zieky, M. J. (1982). Passing scores. Princeton, NJ: Educational Testing Service.

Mathworks, Inc. (1995). MATLAB [computer program]. Englewood Cliffs, NJ: Prentice Hall.

Nedelsky, L. (1954). Absolute grading standards for objective tests. Educational and Psychological Measurement, 14, 3-19.

Overall, J. E., & Klett, C. J. (1972). Applied multivariate statistics. New York: McGraw-Hill.

Peters, E. (1981). Basic skills improvement policy implementation guide #3: Standards setting manual. Boston, MA: Massachusetts Department of Education. (ERIC Document Reproduction Service No. ED 206 696)

Poggio, J. P., Glasnapp, D. R., & Eros, D. S. (1982, March). An evaluation of contrasting groups methods for setting standards. Paper presented at the annual meeting of the American Educational Research Association, New York, NY.

Reid, J. B. (1991). Training judges to generate standard-setting data. Educational Measurement: Issues and Practice, 10(2), 11-14.

Schaeffer, G. A., Carlson, R. E., & Matas, R. L. (1986). Assessing the reliability of criterion-referenced measures used to evaluate health education programs. Educational Review, 10(1), 115-125.

Shepard, L., Glaser, R., Linn, R., and Bohrnstedt, G. (1993). Setting performance standards for student achievement. Stanford, CA: National Academy of Education.

Subkoviak, M. J. (1984). Estimating the reliability of mastery-nonmastery classifications. In R. A. Berk (Ed.) A guide to criterion-referenced test construction (pp. 267-291). Baltimore, MD: Johns Hopkins University Press.

Ziomek, R. L., & Szymczuk, M. (1983). A comparison of approaches for setting proficiency standards via Monte Carlo simulations. Department of Evaluation and Research, DesMoines Public Schools, Iowa. (ERIC Document Reproduction Services No. ED 236 231)

Figure 1

Hypothetical Plot of Master and Nonmaster Distributions

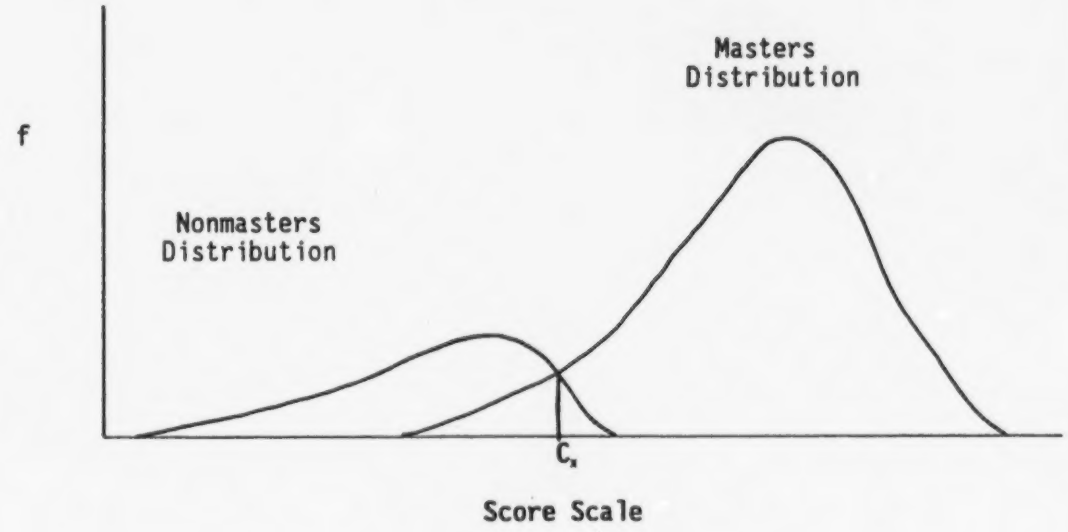
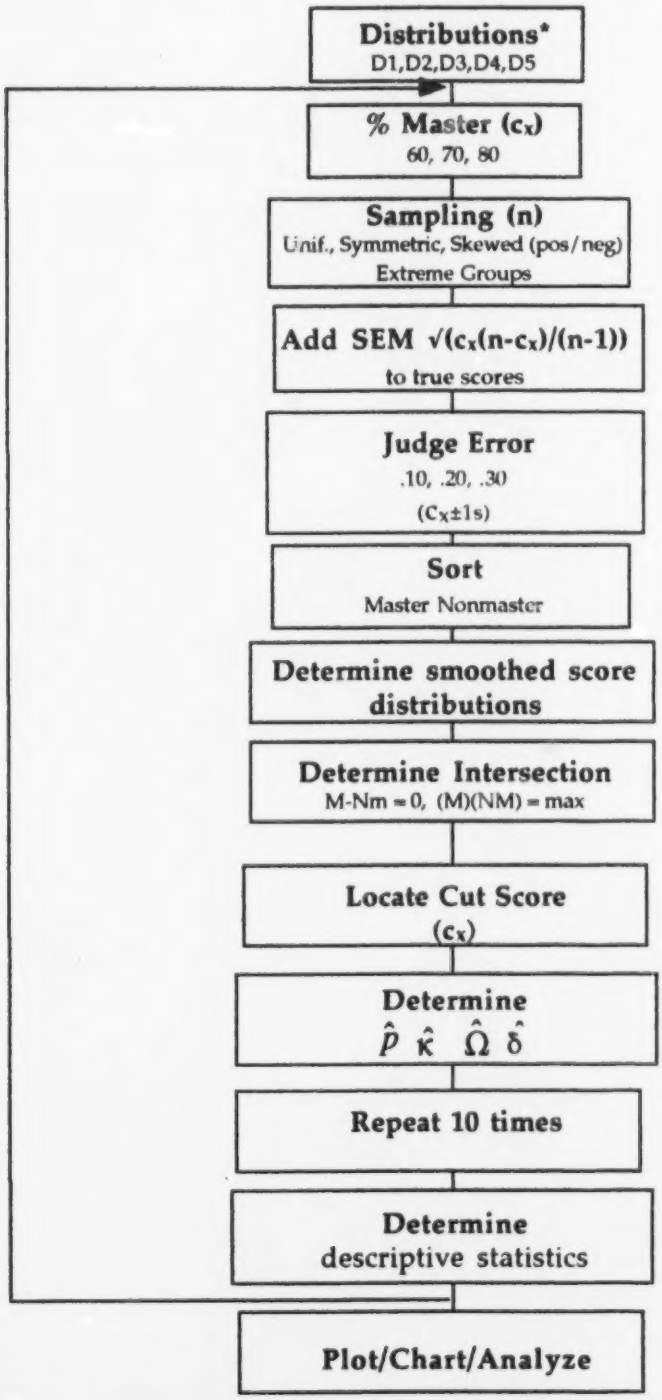


Figure 2

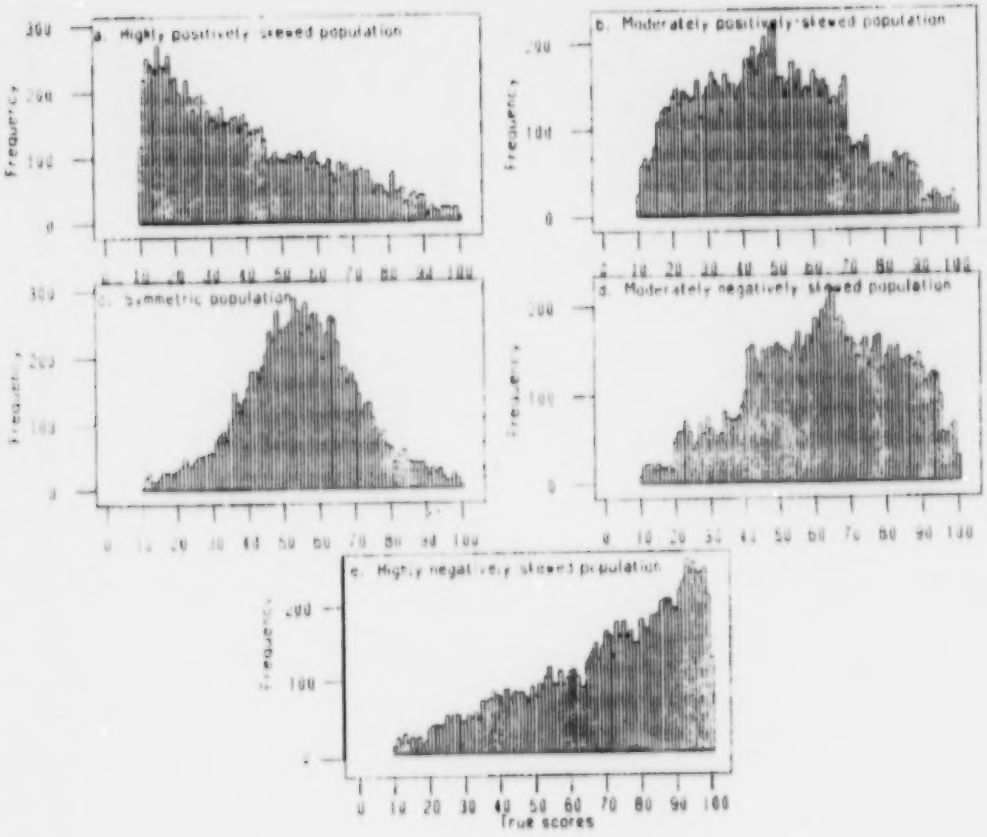
Flow Chart for Simulation Study



*Repeated for samples from these populations for samples of sizes 24, 102, and 1020

Figure 3

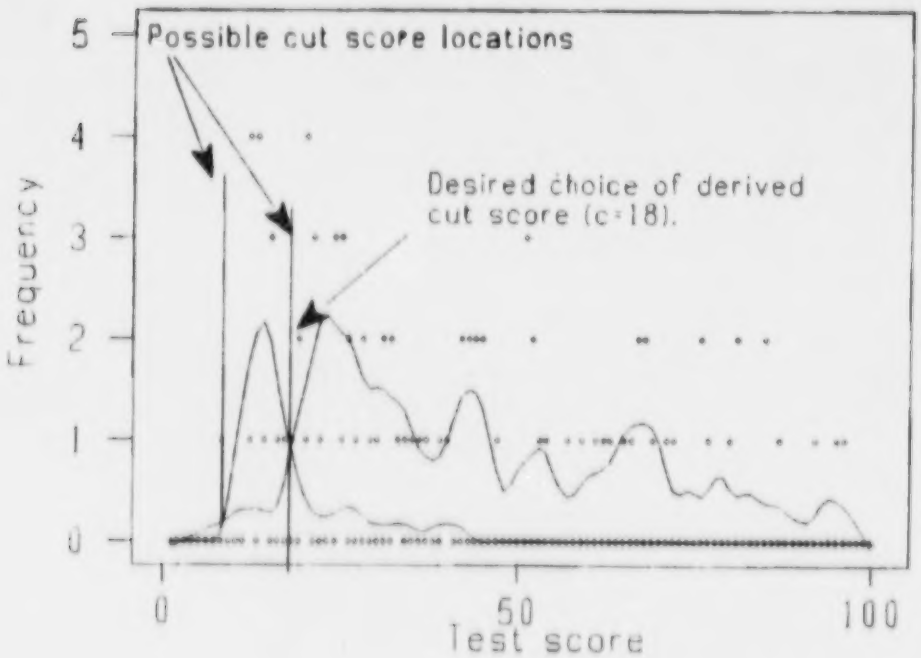
Simulation Study Population Distributions



BEST COPY AVAILABLE

Figure 4

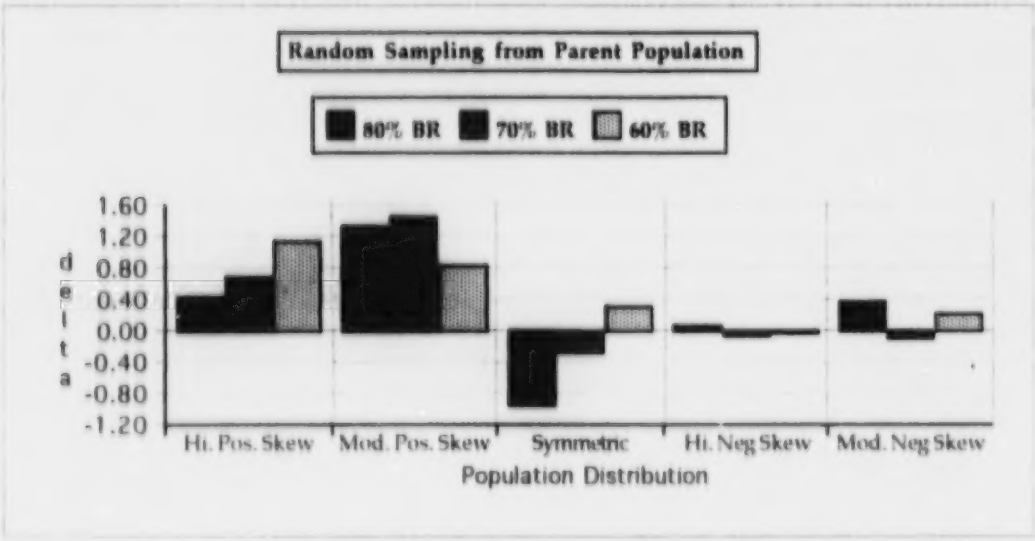
Illustration of Multiple Intersection Rule



BEST COPY AVAILABLE

Figure 5

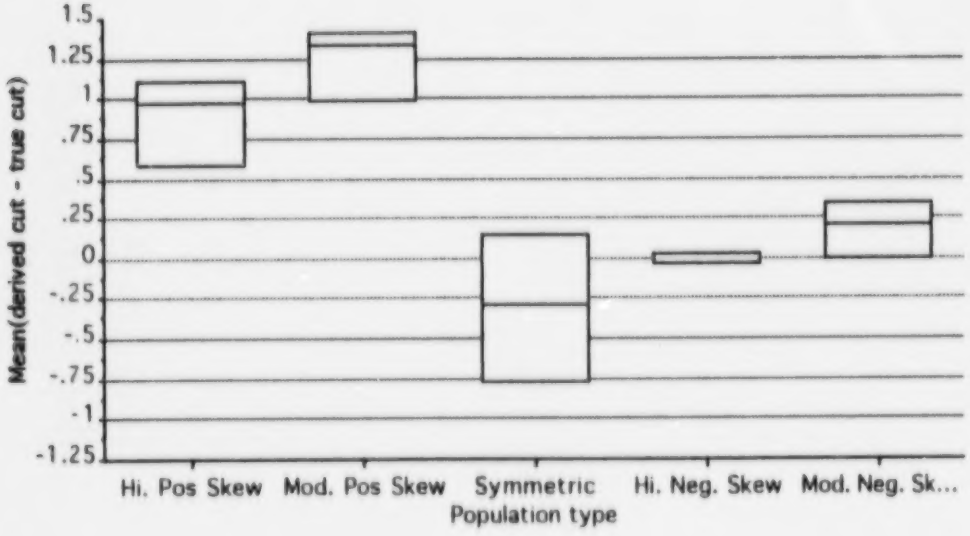
Mean δs for Five Parent Distributions; Sampling Strategy Like the Parent Population; Sample Size = 1020; Judge Error Rate = 30%



BEST COPY AVAILABLE

Figure 6

Box Plots of 10 Mean ds for Five Parent Distributions; Sampling Strategy Like the Parent Population; Sample Size = 1020; Judge Error Rate = 30%



BEST COPY AVAILABLE

Figure 7

Mean δs across Population Base Rates and Sampling Strategies;

Moderately Negatively-Skewed Population Distribution

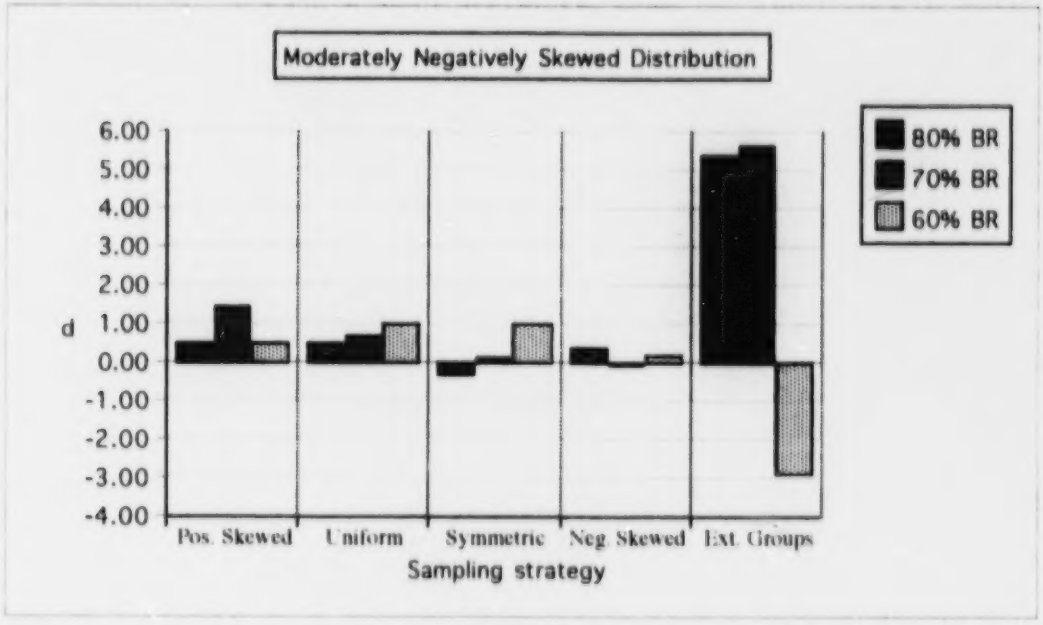
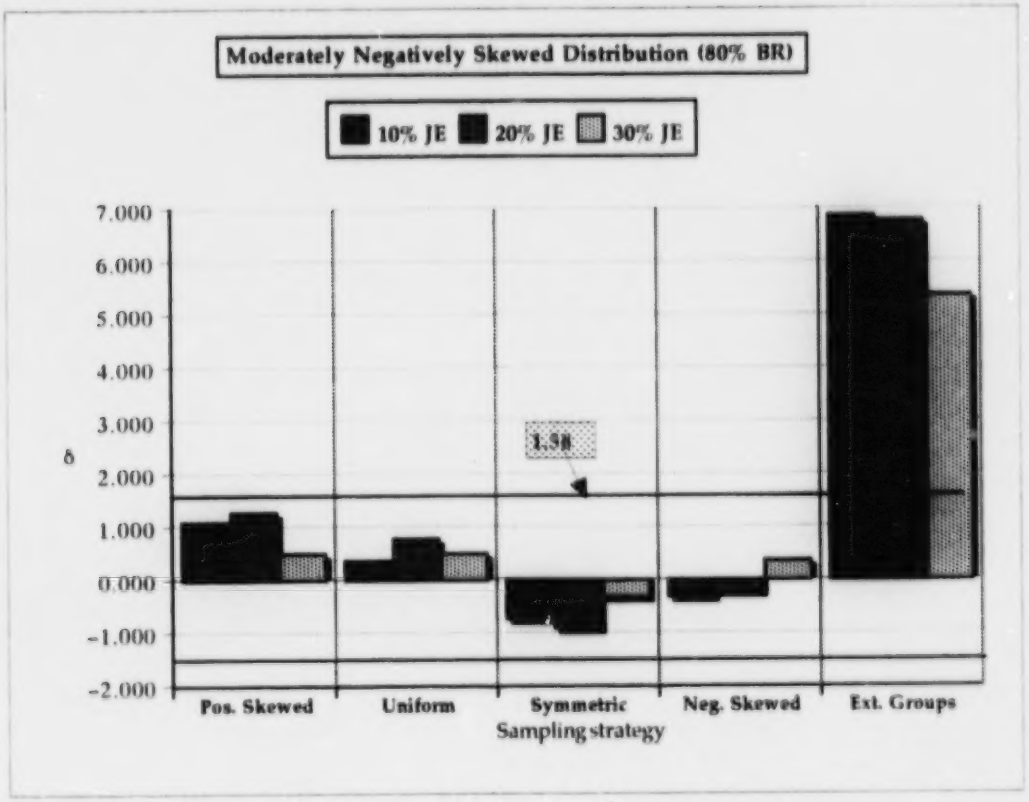


Figure 8

Histogram for δs for Judge Error Rates across Sampling Strategies; Base Rate = 80%;

Moderately Negatively-Skewed Population Distribution; Sample Size = 1020



BEST COPY AVAILABLE

Table 1

Standard Deviations of Derived Cut Scores across Sampling Strategies, Base Rates, and Judge Error Rates; Moderately Negatively-Skewed Population; Sample Size = 1020.

		Negatively Skewed Population			
	Pos. Skewed	Uniform	Symmetric	Neg. Skewed	Extremegroup
Base Rate (80%)					
Judge Error					
0.1	0.73786	1.0328	0.99443	0.84984	4.99
0.2	1.1005	1.07497	0.73786	0.69921	3.5024
0.3	1.25167	1.63639	1.35401	1.39841	3.7059
Base Rate (70%)					
Judge Error					
0.1	0.31623	0.56765	0.82327	0.4714	16.9801
0.2	0.6667	1.68655	1.0328	0.94868	3.4319
0.3	0.99443	1.42984	0.91894	1.50555	4.3218
Base Rate (60%)					
Judge Error					
0.1	0.63246	0.8756	0.56765	1.0328	0.7379
0.2	1.44914	0.99443	0.69921	1.37032	2.5298
0.3	1.3333	1.17851	0.84984	1.70294	5.9479

BEST COPY AVAILABLE

Notes

1. Though not a focus of this paper, the definition, qualifications, and training of "experts" for standard-setting are critical elements in and of themselves. Readers are referred to literature which focusses on this topic, including Cizek (1996), Jaeger (1989b; 1991), and Reid (1991).